Com funcionen els models massius de llengua de l'estil de ChatGPT

Mikel L. Forcada[1,2]

[1]Departament de Llenguatges i Sistemes Informàtics,
Universitat d'Alacant, E-03071 Alacant (fins a 03/02/2024)

[2]Prompsit Language Engineering, S.L.,
Edifici Quorum III, Av. Universitat s/n, E-03202 Elx

Seminari IULATERM, Universitat Pompeu Fabra
16 de gener del 2024

Índex

Índex

Model de llengua

Un model probabilístic de llengua, o simplement model de llengua és un dispositiu matemàtic que assigna una probabilitat (o versemblança) P a una seqüència de mots:

$$P(\textit{El president del govern}) > P(\textit{El president del gat})$$

$$P(\textit{El meu gat és negre}) > P(\textit{El meu negre és gat})$$

$$P(\textit{El meu gat és negre}) > P(\textit{El meu gat és verd})$$

Models de *n*-grames/1

- Els models de llengua basats en *n-grames* calculen la probabilitat P mot a mot.
- Es basen a predir la probabilitat d'un mot després de $n-1$ mots.
- Per exemple, si $n = 3$ (model de "trigrames") les probabilitats són de la forma

$$p_3(\textit{verd} \mid \textit{gat és}),$$

(és a dir, la probabilitat del mot *verd* darrere dels mots *gat és*).

Models de *n*-grames/2

Un model de trigrames calcularia la probabilitat d'*El meu gat és negre* en trams de tres mots:

$$
\begin{aligned}
P(\textit{El meu gat és negre .}) = \quad & p_3(\textit{El} \mid -\,-) \times \\
\times \quad & p_3(\textit{meu} \mid -\,\textit{El}) \times \\
\times \quad & p_3(\textit{gat} \mid \textit{El meu}) \times \\
\times \quad & p_3(\textit{és} \mid \textit{meu gat}) \times \\
\times \quad & p_3(\textit{negre} \mid \textit{gat és}) \\
\times \quad & p_3(. \mid \textit{és negre})
\end{aligned}
$$

Models de *n*-grames/3

- Les probabilitats de l'estil de $p_3(\textit{és} \mid \textit{meu gat})$ s'*estimen*.
- L'estimació es fa *comptant* en un corpus de grans quantitats de text (model estadístic)
- quantes voltes apareix la seqüència *meu gat* ***mot***
- per a tots els possibles mots ***mot***.
- Les probabilitats $p_3(\textit{és} \mid \textit{meu gat})$ es guarden en una *taula*.

Models de *n*-grames/4

- Els models de *n*-grames consideren només un context de $n-1$ mots a l'hora de predir el següent mot.
- Podem augmentar n per a tenir més context, però,
- si el vocabulari té una grandària V,
- necessitaríem una taula amb

$$V^n = \underbrace{V \times V \times V \times \ldots \times V}_{n \text{ voltes}}$$

entrades.

- Si V són 10.000 mots (vocabulari menut) i n és un context (molt menut!) de 4 mots ("tetragrames"), podríem tenir $10.000 \times 10.000 \times 10.000 \times 10.000 =$ 10.000.000.000.000.000 entrades!

Models de *n*-grames/5

- 10.000.000.000.000.000 tetragrames no cabrien en la memòria de la majoria dels ordinadors.
- A més, no tots els 10.000.000.000.000.000 tetragrames ($n = 4$) apareixen en un corpus de text.
- Si en un text nou apareix un tetragrama no vist, hem d'estimar la probabilitat amb un trigrama ($n = 3$);
- si no, amb un bigrama ($n = 2$), etc. $\rightarrow$ pèrdua de context $\rightarrow$ predicció pitjor.

Models de *n*-grames/6

Models massius de llengua:[1]

- En un model massiu de llengua, $n > 1.000$, per exemple, $n = 2.049$.[2]
- Clarament, no funciona amb una taula (no cabria en l'univers!).
- Per tant, no es pot fer comptant esdeveniments en un corpus i guardant les probabilitats.
- Cal un mecanisme per a *calcular*

$$p_{2049}(\textbf{\textit{mot}} \mid \ldots \textit{(els 2.048 mots anteriors)} \ldots)$$

[1] En anglés, *large language models* (LLM).
[2] És a dir, $2^{11} + 1 = 2.048 + 1 = 2.049$

Predir el mot següent no és fàcil /1

- Els models de llengua aprenen a *predir el mot que segueix* els anteriors.
- Aquesta tasca no és gens fàcil.
- Per a fer-la amb èxit, s'han d'aprofitar les pistes que hi ha en els mots anteriors.

Predir el mot següent no és fàcil /2

- Hi ha prediccions fàcils (“apreses de memòria”):

  The former president of the United States, Donald[3] $\rightarrow$

[3] *L’anterior president dels Estats Units, Donald* $\rightarrow$

Predir el mot següent no és fàcil /2

- Hi ha prediccions fàcils ("apreses de memòria"):

  The former president of the United States, Donald[3] → Trump

[3] *L'anterior president dels Estats Units, Donald* →

Predir el mot següent no és fàcil /3

- Hi ha prediccions on s'ha de fer un processament ("raonament") més complex:

  As the king only had a daughter, the next monarch, instead of a king, would be a[4] $\rightarrow$

[4] *Com que el rei només tenia una filla, el monarca, en comptes d'un rei, seria una* $\rightarrow$

Predir el mot següent no és fàcil /3

- Hi ha prediccions on s'ha de fer un processament ("raonament") més complex:

  As the king only had a daughter, the next monarch, instead of a king, would be a[4] $\rightarrow$ queen

(prediccions correctes d'un model de llengua relativament menut, `EleutherAI/gpt-neo-1.3B`)

[4] *Com que el rei només tenia una filla, el monarca, en comptes d'un rei, seria una* $\rightarrow$

Predir el mot següent no és fàcil /4

Assumim que els models de llengua no *entenen*. Però, sense "entendre" es pot predir el mot següent?

Provem-ho!

Predir el mot següent no és fàcil /4

Assumim que els models de llengua no *entenen*. Però, sense "entendre" es pot predir el mot següent?

Provem-ho!

Un fretidor de rebulació ímpola és una rèdula que transforma en una cítula de rebulació l'ocalígia víntrica d'un rebulable en ocalígia $\rightarrow$?????

No sabem ni de què ens parlen. No podem predir el mot següent.

Predir el mot següent no és fàcil /5

Però els models de llengua s'*entrenen* amb grans quantitats de text, que podrien contenir, per exemple:

> *[. . .] Un fretidor de rebulació ímpola és un tipus de rèdula que obté ocalígia rèdia directament de l'ocalígia víntrica produïda per un rebulable que rebula dins d'una cítula de rebulació, la part principal d'un fretidor. [. . .]*
> *Una classe important de rèdula és el fretidor de rebulació ímpola, sovint anomenat simplement fretidor de rebulació. En la cítula de rebulació d'aquest tipus de fretidor es rebula un rebulable per a generar ocalígia rèdia a partir de la seua ocalígia víntrica. [. . .]*

Predir el mot següent no és fàcil /6

No entenem res, però podem predir ara el mot següent?

Predir el mot següent no és fàcil /6

No entenem res, però podem predir ara el mot següent?

Provem una altra volta:

Predir el mot següent no és fàcil /6

No entenem res, però podem predir ara el mot següent?

Provem una altra volta:

Un fretidor de rebulació ímpola és una rèdula que transforma en una cítula de rebulació l'ocalígia víntrica d'un rebulable en ocalígia $\rightarrow$?????

Podem tornar a rellegir el text anterior.

Predir el mot següent no és fàcil /7

[. . .] Un fretidor de rebulació ímpola és un tipus de rèdula que obté ocalígia rèdia directament de l'ocalígia víntrica produïda per un rebulable que rebula dins d'una cítula de rebulació, la part principal d'un fretidor. [. . .]
Una classe important de rèdula és el fretidor de rebulació ímpola, sovint anomenat simplement fretidor de rebulació. En la cítula de rebulació d'aquest tipus de fretidor es rebula un rebulable per a generar ocalígia rèdia a partir de la seua ocalígia víntrica. [. . .]

Predir el mot següent no és fàcil /8

I ara, podem predir el mot següent?

Un fretidor de rebulació ímpola és una rèdula que transforma en una cítula de rebulació l'ocalígia víntrica d'un rebulable en ocalígia $\rightarrow$

Predir el mot següent no és fàcil /8

I ara, podem predir el mot següent?

Un fretidor de rebulació ímpola és una rèdula que transforma en una cítula de rebulació l'ocalígia víntrica d'un rebulable en ocalígia $\rightarrow$ rèdia

Predir el mot següent no és fàcil /9

Els exemples eren textos reals amb alguns mots emmascarats.

Predir el mot següent no és fàcil /10

El context:
[...] Un motor de combustió interna és un tipus de màquina que obté energia mecànica directament de l'energia química produïda per un combustible que crema dins d'una cambra de combustió, la part principal d'un motor. [...]
Una classe important de màquina és el motor de combustió interna, sovint anomenat simplement motor de combustió. En la cambra de combustió d'aquest tipus d'emmmetidor es crema un combustible per a generar energia mecànica a partir de la seua energia química. [...]

Predir el mot següent no és fàcil /11

El problema:

Un motor de combustió interna és una màquina que transforma en una cambra de combustió l'energia química d'un combustible en energia $\rightarrow$

Predir el mot següent no és fàcil /11

El problema:

Un motor de combustió interna és una màquina que transforma en una cambra de combustió l'energia química d'un combustible en energia $\rightarrow$ mecànica.

Predir el mot següent no és fàcil /12

Què ha passat? Per a poder predir el mot següent, hem hagut de fer mentalment algunes *manipulacions abstractes* amb mots que no havíem sentit mai.

- Hem provat d'entendre les relacions que hi havia entre els mots.
- Hem intentat crear representacions abstractes dels mots en el nostre cap.
- Hem usat les relacions entre aquestes representacions per a predir el mot següent.[5]

[5] Per cert, que hem demanat a Google Bard i ChatGPT "Llegiu els primers dos paràgrafs atentament, i després completeu el tercer amb un únic mot.": el primer ho ha fet bé i el segon, malament.

Predir el mot següent no és fàcil /13

El repte: com calcular

$$p_{2049}(\textbf{\textit{mot}} \mid \ldots \textit{(els 2.048 mots anteriors)} \ldots)?$$

Recordeu:

- No és factible guardar els 2.049-grames en una taula (si el vocabulari fora de 10.000 mots tindríem una taula amb 10000^{2049} entrades, és a dir, un 1 seguit de 8.196 zeros.
- No hi ha text en l'univers on es puguen trobar tots els 2.049-grames.

Per tant, quan el model veu 2.048 mots d'un text i li demanem el 2.049é, no pot recordar-se'n. L'ha de *deduir*, una miqueta com hem fet nosaltres.

Índex

Neurones artificials /1

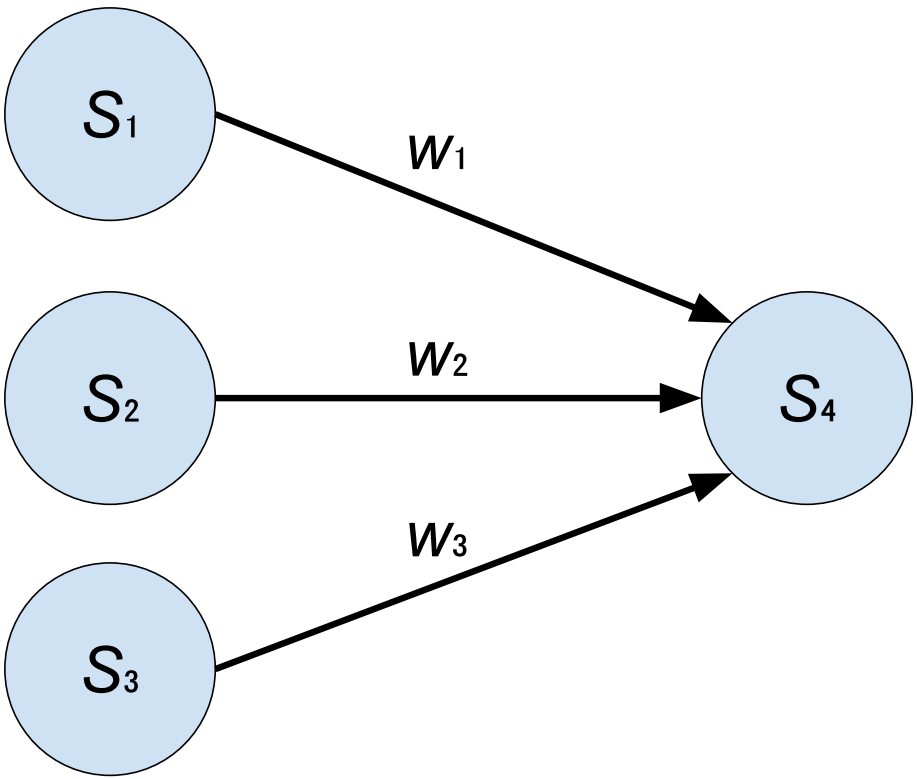
Per què es diu que els models de llengua massius són neurals (o neuronals)?

- Perquè són grans xarxes de *neurones artificials*, *simulades* per ordinador (no construïdes físicament).
- L'*activació* (excitació) de cada neurona depén de l'activació d'altres neurones i de les característiques de les connexions mútues.
- El signe i la magnitud dels *pesos* d'aquestes connexions determinen el comportament de la xarxa:
  - Les neurones connectades amb pes positiu tendeixen a excitar-se o inhibir-se simultàniament.
  - Les neurones connectades amb pes negatiu tendeixen a ser en estats oposats.
  - L'efecte de la interacció augmenta amb la magnitud del pes.

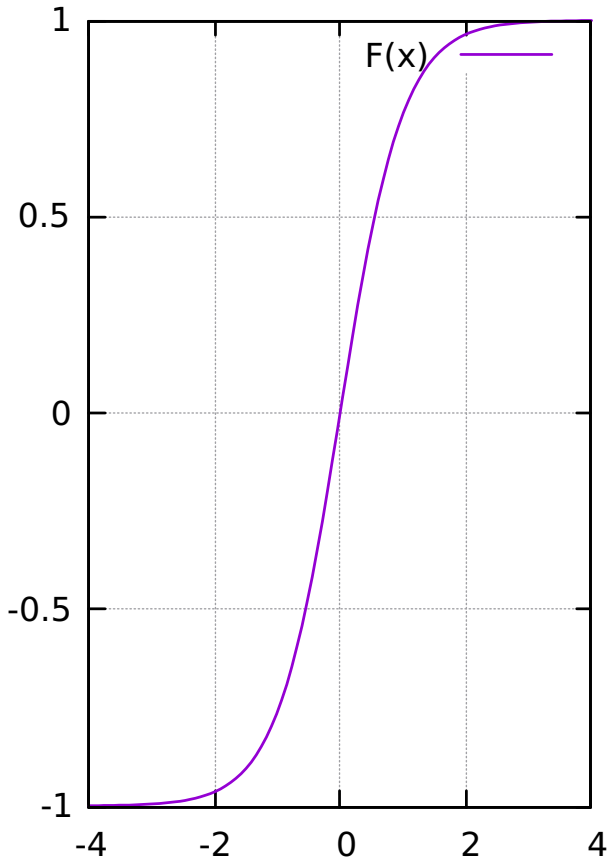
Neurones artificials /2

- L'*entrenament* de la xarxa neural artificial fixa els pesos als valors necessaris per a assegurar un comportament determinat:
  - $\rightarrow$ determinats patrons d'excitació o d'inhibició.

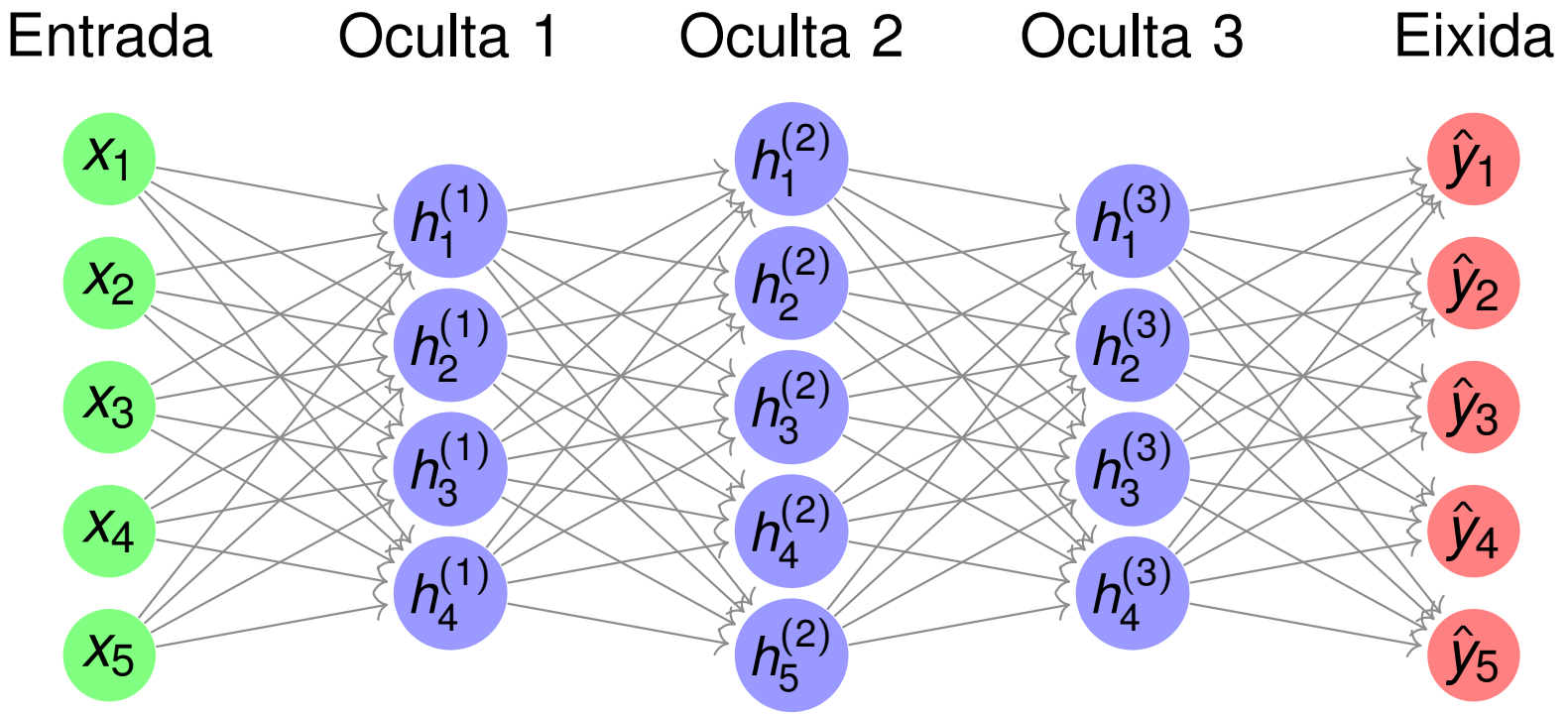
Neurones artificials /3



$$S_4 = F\,(w_1 \times S_1 + w_2 \times S_2 + w_3 \times S_3)$$



Xarxes neurals



Una xarxa neural multicapa amb 5 entrades, 4 neurones en la 1a capa oculta, 5 neurones en la 2a capa oculta, 4 neurones en la 3a capa oculta i 5 neurones d'eixida.

Representacions /1

Quan les xarxes neurals s'usen per a fer tasques lingüístiques:

- S'assigna un número a cada mot (o submot) del vocabulari.
  - Per exemple, s'assigna el número 34 al mot *casa*.
- Totes les entrades s'apaguen (senyal=0,0) excepte la corresponent al mot o submot, que s'encén (senyal=1,0).
- La grandària del vocabulari d'entrada està limitada a un nombre de mots (o submots): tantes com entrades.
- Les eixides s'interpreten com a probabilitats.
  - L'activació de la neurona d'eixida número k és proporcional a la probabilitat del k-ésim mot (o submot) del vocabulari.

Representacions /2

- Quan s'entrenen amb alguna tasca lingüística, els valors d'activació de les capes de neurones formen *representacions* de la informació que processen.
- Per exemple, el *vector*

$$(0{,}35, 0{,}28, -0{,}15, 0{,}76, \ldots, 0{,}88)$$

podria ser la representació del mot *estudi*, i el *vector*

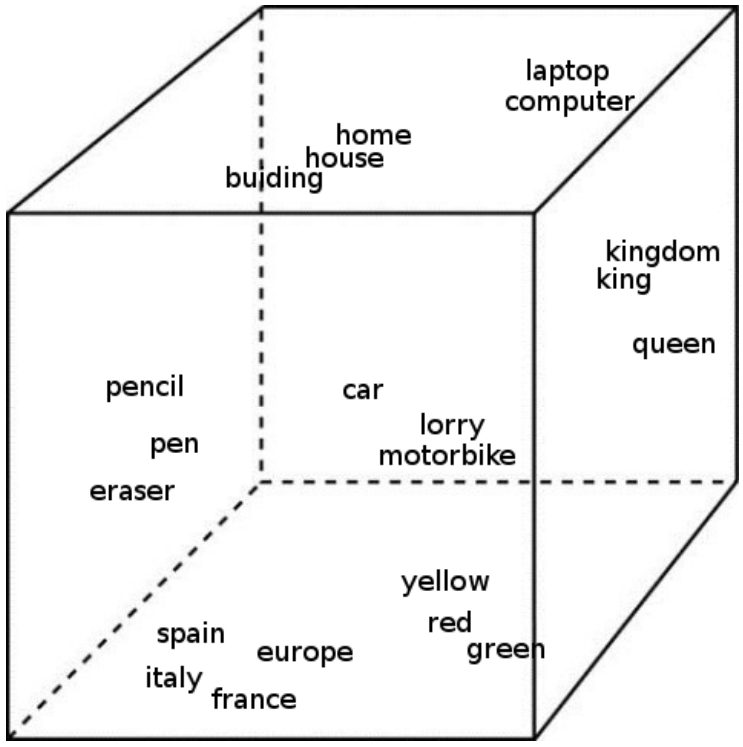
$$(0{,}93, -0{,}78, 0{,}22, 0{,}31, \ldots, -0{,}71)$$

la del mot *gat*.

- Diem que els mots *es representen* en un *espai* (vectorial).

Representacions /3

Imaginem representacions lèxiques amb només tres neurones (no és fàcil imaginar espais de més dimensions):



Els mots amb interpretacions similars són a prop uns dels altres.

Representacions /4

- Les representacions apreses solen ser tals que s'hi pot fer aritmètica semàntica (afegint i substraent valors d'activació neurona a neurona). Si la representació de cada mot és una columna en un full de càlcul, i operem amb les columnes, observarem que

$$[\text{rei}] - [\text{home}] + [\text{dona}] \simeq [\text{reina}]$$

- Això indica que la xarxa prova de representar les relacions entre paraules (la "semàntica").

Índex

Models neurals de llengua/1

Els models neurals de llengua

- són *predictors*: prediuen els mots un a un.
  - És paregut a com el vostre telèfon mòbil prediu el mot següent quan teclegeu un missatge.
- Més precisament, *estimen la probabilitat* o la *versemblança* $p(t)$, per a cada possible mot (o submot) t del vocabulari V, que aquest mot (o submot) haja d'aparéixer en la posició actual.

Models neurals de llengua/2

En els models neurals de llengua:

- El procés de *predicció* selecciona el(s) (sub)mot(s) més probable(s).
- La probabilitat del (sub)mot t en la posició k de l'eixida, $p_k(t)$, depén de la seqüència dels M (sub)mots d'eixida anteriors

$$t_{k-M}\ t_{k-M+1}\ t_{k+M-2}\ \ldots\ t_{k-2}\ t_{k-1};$$

- matemàticament:

$$p_k(t \mid t_{k-M}\ t_{k-M+1}\ t_{k+M-2}\ \ldots\ t_{k-2}\ t_{k-1})$$

Submots?

Els models de llengua treballen amb unitats lèxiques (anglés *tokens*) que poden ser mots o "submots":

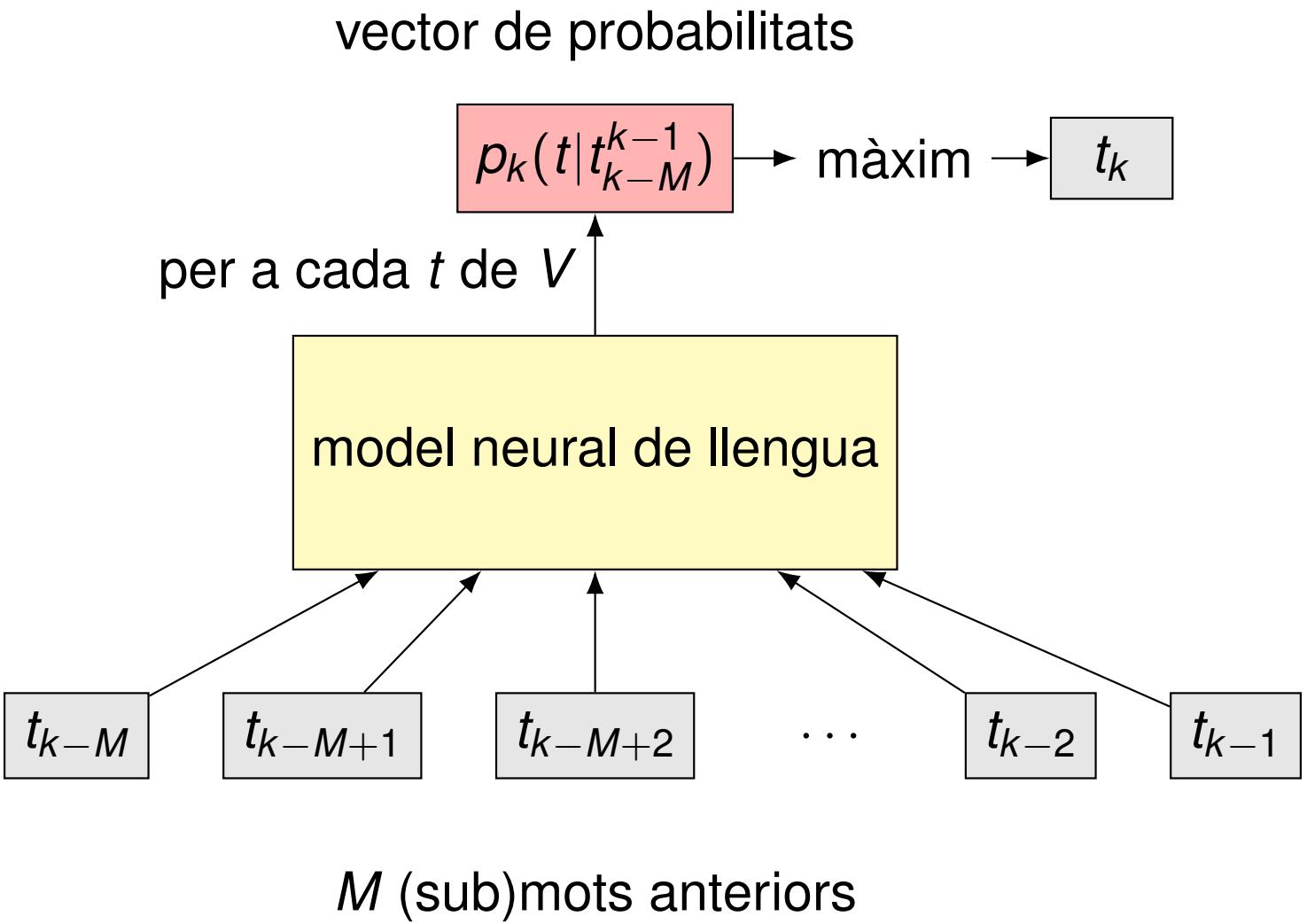
- Per als mots freqüents, el segment és el mot
- Per als mots menys freqüents, el mot es pot partir en segments més menuts:

  Passe- j- à- ve- m per les ext- ense- s a- vi- ngu- des .

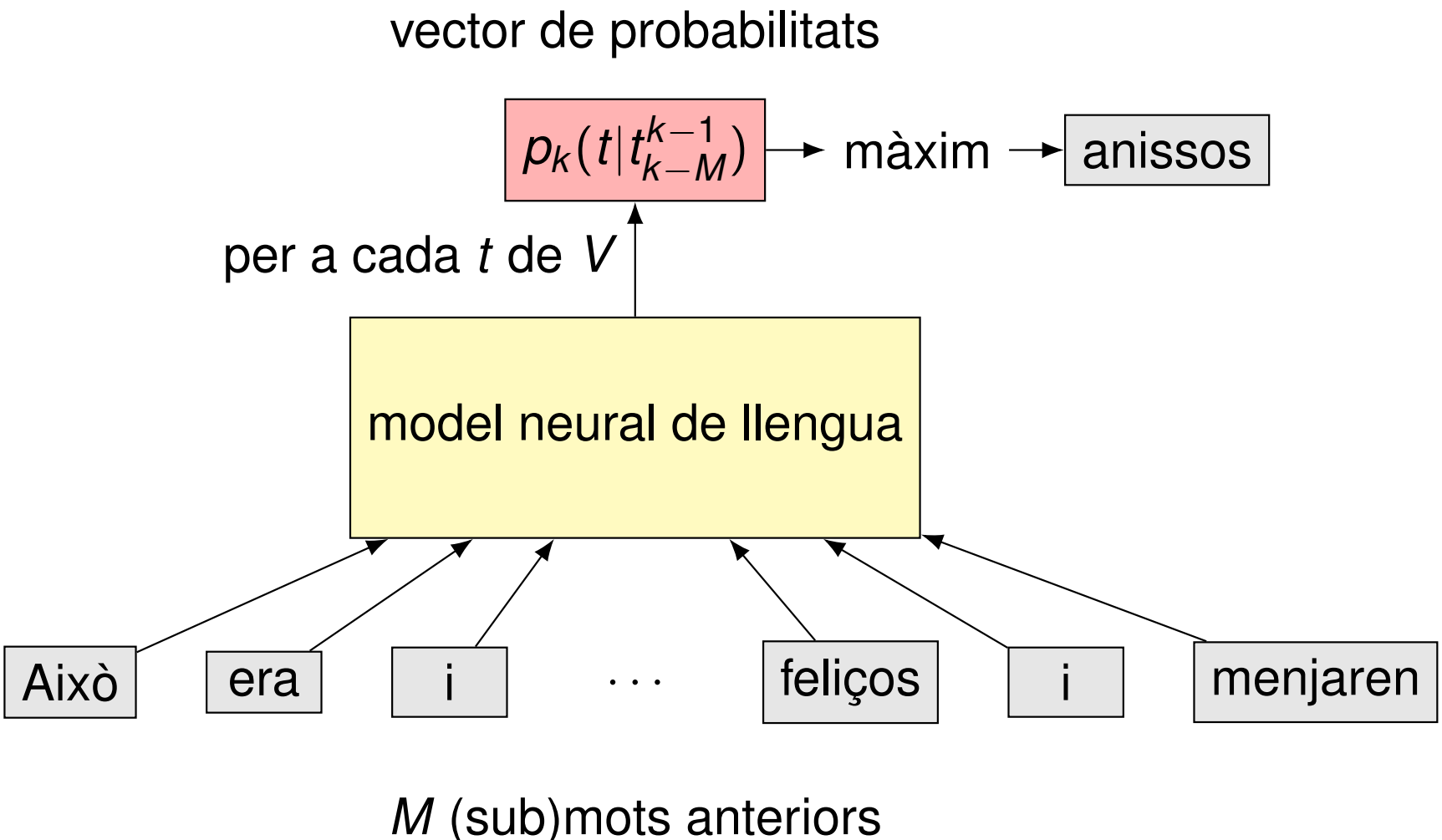
- Aquesta segmentació s'aprén automàticament a partir del corpus de manera no supervisada (*byte-pair encoding*, *SentencePiece*, etc.).

Així s'eviten xarxes immenses, ja que la xarxa té una entrada i una eixida per a cada (sub)mot del vocabulari.

El model de llengua com una caixa negra /1



El model de llengua com una caixa negra /2



El model de llengua com una caixa negra /3

En un model de llengua ben entrenat amb contes infantils,

$$p_k(\text{“anissos”}|\text{“Això era i... feliços i menjaren”})$$

hauria de ser més alta que

$$p_k(\text{“elefants”}|\text{“Això era i... feliços i menjaren”})$$

De fet, idealment, hauria de ser *la més alta* per a tots els (sub)mots en V:

$$p_k(\text{“anissos”}|\text{“Això era i... feliços i menjaren”}) > p_k(t|\text{“Això era i... feliços i menjaren”})$$

per a tot altre t del vocabulari.

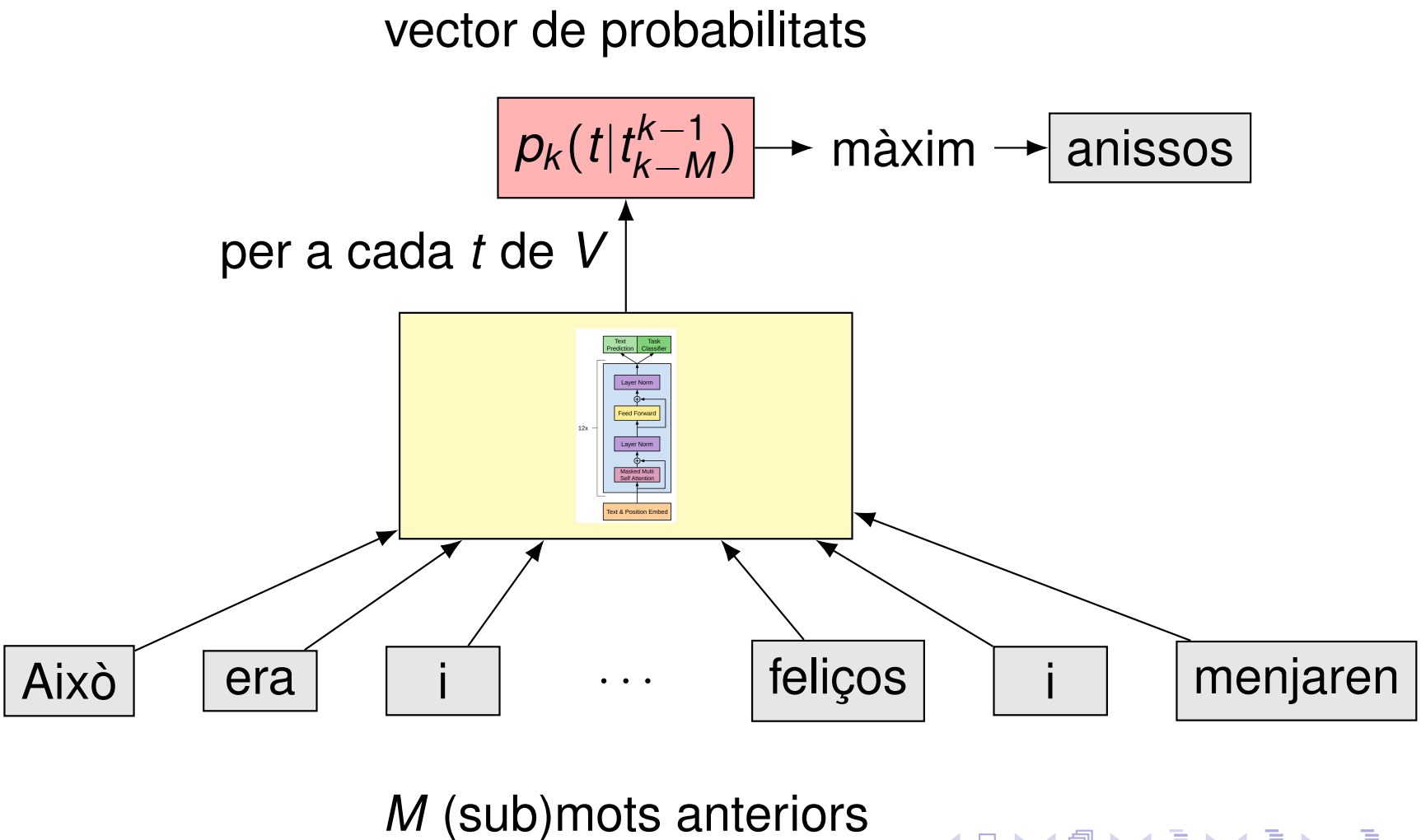
Els *transformers*

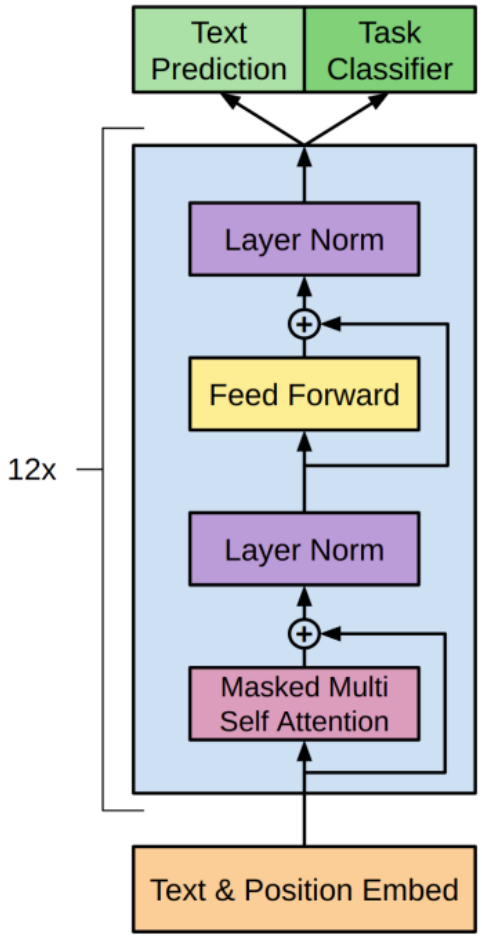
Els càlculs que es fan en la caixa groga es poden fer amb *arquitectures* neurals ben diverses, però la més comuna són els *transformers*.[6]

- No és fàcil d'explicar, però ho intentaré.
- Són 11 diapositives *només*.

[6] Vaswani, N., et al. (2017) "Attention is all you need". In: *Advances in Neural Information Processing Systems*, pp. 6000–6010.

Un *transformer* dins de la caixa negra





Radford et al. (2018) “Improving Language Understanding by Generative Pre-Training” (informe intern d’OpenAI.)

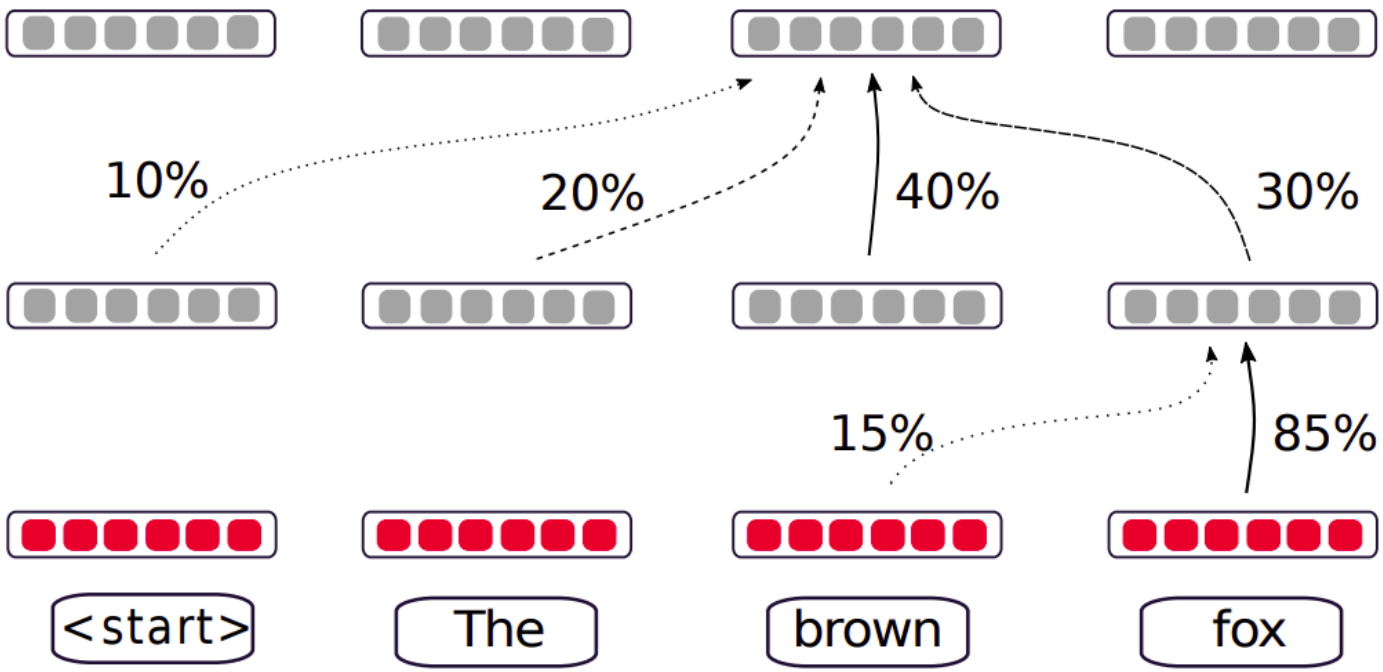
Els *transformers*

Un *transformer* funciona amb mecanismes d'*atenció*, i ho fa per etapes (cada una, basada en múltiples capes):

- En cada etapa, incorpora context a la representació de cada (sub)mot anterior *parant atenció selectiva* a les representacions de tots els (sub)mots de l'etapa anterior.
- En la última etapa, es presta atenció a totes les representacions per a generar l'eixida:
  - Probabilitats de cada mot en la posició següent (predicció)
  - Una altra tasca (ara ho veurem).
- El nombre d'etapes és típicament 12.
- Les representacions són vectors de 2048 components.[7]

[7] Un nombre redó en informàtica, ja que $2^{11} = 2048$.

Els *transformers*



[8] Figura treta de Pérez-Ortiz, J.A., Forcada, M.L., Sánchez-Martínez, F., "How neural machine translation works", in Kenny, D., ed., Machine translation for everyone: Empowering users in the age of artificial intelligence., 141–164

L'atenció/1

Metàfora per a evitar matemàtiques:[9]

- Cada etapa és una generació en la història d'una població.
- Cada representació en una etapa és un individu en una generació.
- Les representacions (individus) d'una etapa (generació) usen una espècie de *xarxa social de cites* i després *s'aparellen i es reprodueixen* per a generar les representacions de l'etapa (generació) següent.

[9] Inspirada en `https://youtu.be/aL-EmKuB078`, video de DotCSV.

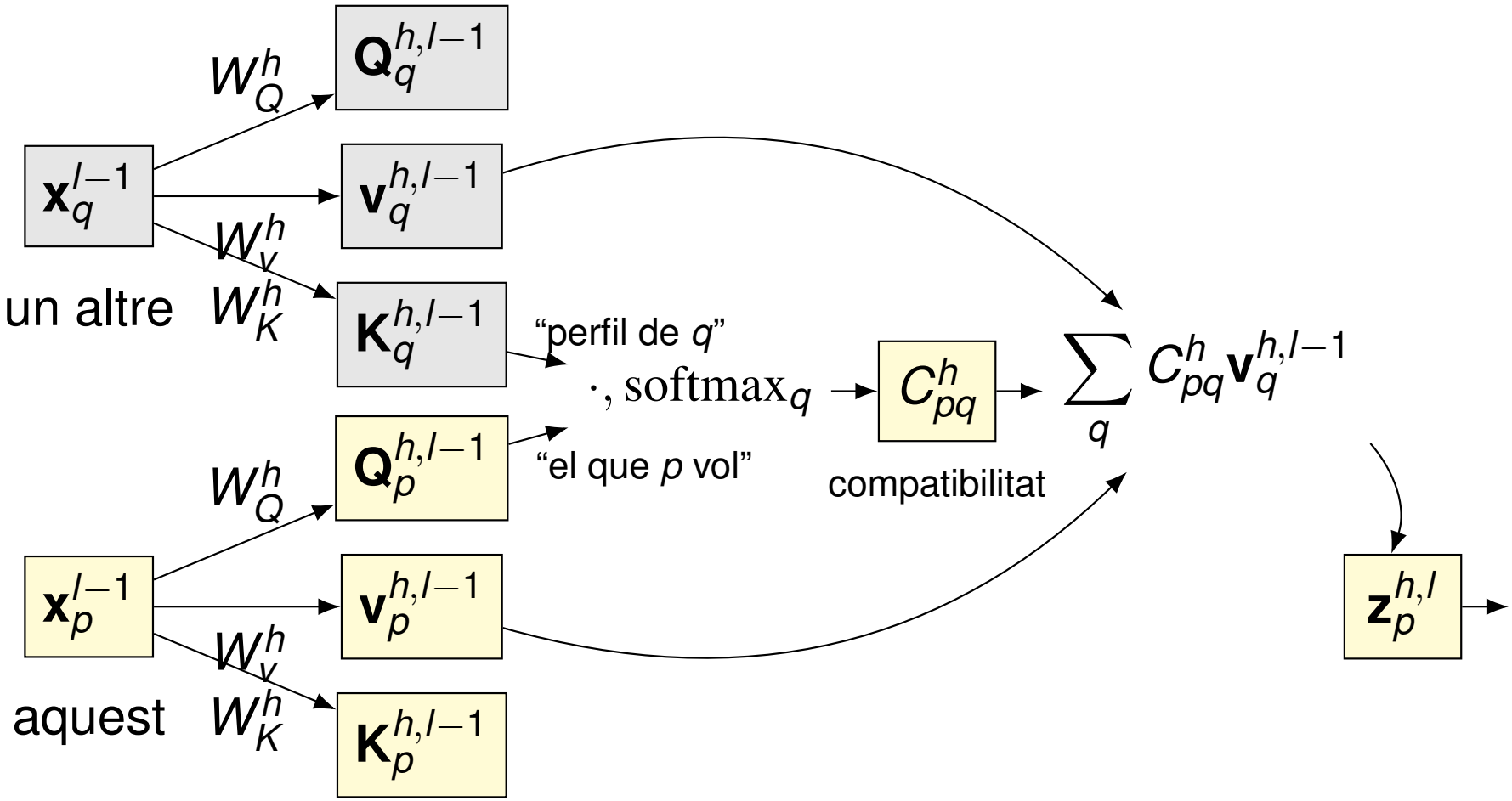
L'atenció /2

Fase d'aparellament:

- Cada representació (individu) en cada etapa:
  - ofereix un perfil a la resta de representacions (oferta, vector *key*);
  - genera un perfil que representa amb qui es volen aparellar (demanda, vector *query*);
  - genera un material genètic que es combinarà (vector *value*);

  (tres vectors)
- A partir de la demanda d'una representació i l'oferta de les altres representacions es genera un percentatge de compatibilitat (atenció).

L'atenció /2'


W_Q^h
$\mathbf{Q}_q^{h,l-1}$
$\mathbf{x}_q^{l-1}$
$\mathbf{v}_q^{h,l-1}$
W_V^h
un altre
W_K^h
$\mathbf{K}_q^{h,l-1}$
"perfil de q"
$\cdot$, softmax$_q$
C_{pq}^h
$\sum_q C_{pq}^h \mathbf{v}_q^{h,l-1}$
"el que p vol"
compatibilitat
W_Q^h
$\mathbf{Q}_p^{h,l-1}$
$\mathbf{x}_p^{l-1}$
$\mathbf{v}_p^{h,l-1}$
W_V^h
aquest
W_K^h
$\mathbf{K}_p^{h,l-1}$
$\mathbf{z}_p^{h,l}$

L'atenció/3

Fase de reproducció:

- El material genètic que la representació passa a l'etapa (generació) següent
  - és una mescla del seu material genètic i el d'altres representacions;
  - el percentatge de mescla (aparellament) és proporcional al percentatge de compatibilitat.
- Les representacions de l'etapa (generació) següent es generen a partir d'aquest material genètic.

L'atenció/4

Per a generar

- les ofertes, les demandes i el material genètic a partir de cada representació (individu) de la generació actual,
- el percentatge de compatibilitat (atenció) a partir de les ofertes i les demandes,
- les representacions (individus) de l'etapa (representació) següent a partir del material genètic.

s'usen xarxes neurals multicapa que s'entrenen perquè el *transformer*

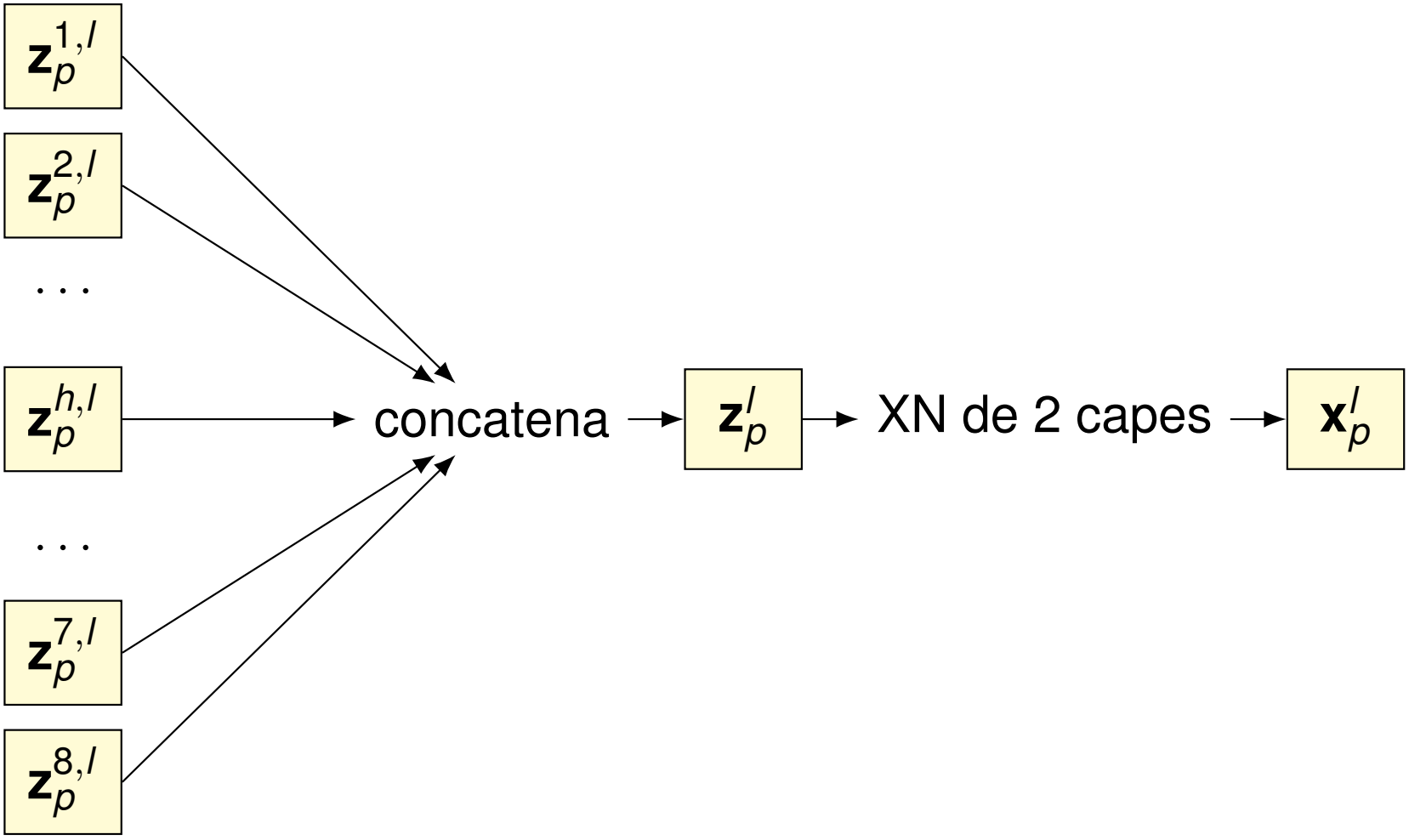
- produïsca les eixides desitjades (probabilitats de cada (sub)mot en la posició següent)
- a partir de les entrades (M (sub)mots anteriors).

L'atenció/5

Realment tot és una miqueta més complicat:

- El procés descrit es fa típicament en 8 fases (*heads* en l'argot dels *transformers*) en cada capa.
- Les representacions inicials dels M mots anteriors duen una *marca d'aigua* que indica la seua posició.
- Hi ha connexions directes addicionals entre capes.

L'atenció /5'


$\mathbf{z}_p^{1,l}$
$\mathbf{z}_p^{2,l}$
. . .
$\mathbf{z}_p^{h,l}$
. . .
$\mathbf{z}_p^{7,l}$
$\mathbf{z}_p^{8,l}$
concatena
$\mathbf{z}_p^{l}$
XN de 2 capes
$\mathbf{x}_p^{l}$

Models massius de llengua: requisits

Entrenar un model massiu de llengua no és a l'abast de tothom:

- Cal obtenir, gestionar, netejar i emmagatzemar un corpus d'entrenament amb milers de milions de mots de text.
- Cal disposar d'un equip informàtic d'alt rendiment equipat amb GPUs (*graphic processing units*).
  - Les CPU usuals de portàtils i ordinadors de taula són massa lentes.
  - L'entrenament neural requereix moltes operacions amb vectors, matrius i tensors,
  - que s'executen molt eficientment en GPU.[10]
- Els *models* es poden *optimitzar* per a executar-se en ordinadors ordinaris.

[10] Per exemple, una GPU A100 amb 80 GB costa $\simeq$ 20,000 EUR i gasta 300 W

Com s'entrenen els models de llengua /1

Entrenar és ajustar tots els pesos de les connexions perquè la xarxa execute una tasca.

Fase inicial: *preentrenament no supervisat*

- Tasca: predicció del mot següent ($\leftarrow$ model de llengua)
- Objectiu: per a cada bloc de M mots, maximitzar la probabilitat que el model assigna al mot que els segueix en el corpus.
- Cal maximitzar aquesta probabilitat per a totes les posicions del corpus de text d'entrenament.

Sovint s'usa el nom *paràmetres* per a referir-se als pesos d'una xarxa neural.

Com s'entrenen els models de llengua /2

- Es comença amb pesos aleatoris.
- L'algorisme calcula quant canvien les probabilitats dels mots predits quan s'incrementa una miqueta cada pes de la xarxa.
- Per a cada pes:
  - (Pes $\uparrow \Rightarrow$ Probabilitat $\uparrow$)? $\Rightarrow$ pes $\uparrow$ proporcionalment.
  - (Pes $\uparrow \Rightarrow$ Probabilitat $\downarrow$)? $\Rightarrow$ pes $\downarrow$ proporcionalment.

Com s'entrenen els models de llengua /3

-
  - (Pes $\uparrow \Rightarrow$ Probabilitat $\uparrow$)? $\Rightarrow$ pes $\uparrow$ proporcionalment.
  - (Pes $\uparrow \Rightarrow$ Probabilitat $\downarrow$)? $\Rightarrow$ pes $\downarrow$ proporcionalment.
- El procés es repeteix fins que
  - passa un temps, o
  - s'ha fet un nombre determinat de passades per tot el corpus, o
  - el comportament en una part reservada del corpus que no hem usat per a entrenar empitjora.

Com s'entrenen els models de llengua /4

- L'algorisme d'aprenentatge calcula un gradient de les probabilitats P predites per a cada mot del text respecte de cada pes w que connecta neurones:

$$\text{gradient}(P, w) = \frac{P(\text{amb } w + \Delta w) - P(\text{amb } w)}{\Delta w} \left(= \frac{\partial P}{\partial w} \right)$$

  - És a dir, quant varia la probabilitat per a un petit canvi Δw en cada pes w.
- Després de processar un cert nombre d'exemples, els pesos s'actualitzen *proporcionalment* a l'efecte sobre la probabilitat $\rightarrow$ *ascens pel gradient*

$$\text{nou } w = w + (\text{taxa d'aprenentatge}) \times \text{gradient}(P, w)$$

- Això es fa repetidament.

Com s'entrenen els models de llengua /5

Segona fase: *Tasques supervisades*: assignar una etiqueta o classe als *M* mots d'entrada.

- El conjunt d'aprenentatge és un corpus de blocs de text etiquetats:
  - Cada bloc de text s'*anota* amb una *etiqueta*.
- S'afig una eixida al transformer amb una neurona per etiqueta o classe.
- Cada tasca té el seu conjunt d'etiquetes o classes.
- Es *reentrena* ("*fine tuning*") el model perquè reproduïsca les classes del corpus d'aprenentatge.

Com s'entrenen els models de llengua /6

Tasques (Radford et al., 2018):

Implicació: Premisa + Hipòtesi $\rightarrow$ V/F

Similitud: Text 1 + Text 2 $\rightarrow$ grau de similitud

Qüestions: Context + Qüestió + Resposta $\rightarrow$ correcte?

Gramaticalitat: Text $\rightarrow$ correcte gramaticalment?

S'usen corpus ja disponibles, creats en investigacions anteriors per a aquestes tasques.[11]

[11] Per exemple, GLUE (*General Language Understanding Evaluation*, `gluebenchmark.com`).

Com s'entrenen els models de llengua /7

Tercera fase: *Retroalimentació humana* (aprenentatge per reforç amb retroalimentació humana).[12]

- Amb un corpus de peticions i respostes correctes, entrenament supervisat del model de llengua (ML) perquè les reproduïsca.
- Per cada petició (d'un altre corpus) es genera una mostra de *N* respostes aleatòries probables però diferents.
- Es demana a informants que les ordenen per ordre de valor (utilitat, acceptabilitat, etc.).

[12] Ouyang et al. (2022) "Training language models to follow instructions with human feedback", informe tècnic d'OpenAI

Com s'entrenen els models de llengua /8

- S'afig una capa al ML que, donades una petició i una resposta, genera una puntuació (model de recompensa, MR).
- S'entrena el MR perquè les puntuacions generades reproduïsquen l'ordenació dels informants.
- Amb noves peticions, es generen respostes, i s'usa l'eixida del MR per a reentrenar ("refinar") el ML.

Índex

Alguns models de llengua existents/1

Hi ha molts models de llengua:

- Alguns són *oberts i lliures*
  - El corpus d'entrenament és públic.
  - L'estratègia i les condicions de l'entrenament són públics.
  - Es pot descarregar el model resultant per a executar-lo localment.
  - Es poden usar lliurement sense restriccions.
- Altres no ho són.
  - Es poden usar gratuïtament o pagant.
  - No es poden descarregar per a executar-los localment.
  - Es poden descarregar però hi ha restriccions d'ús.
  - No se sap com s'han entrenat i sobre quin(s) corpus.

Alguns models de llengua existents /2

Exemples:

- Oberts i lliures:
  - Els models d'EleutherAI[13] (GPT-Neo, GPT-J, GPT-JT; el més gran té 20.000 milions de pesos, entrenats sobre un corpus públic anomenat *The Pile* (886 GB de text).
  - Altres models derivats de GPT-2 com ara GPT4All (assistent virtual)
  - El projecte RedPajama (30.000.000.000.000 submots) vol reproduir models com LLaMA (veure més avall).
- No oberts / no lliures:
  - LLaMA 2 de Meta (7.000–70.000 milions de paràmetres, jul. 2023): corpus no disponibles.
  - ChatGPT 3.5 (el famós ChatGPT: entrenament parcialment desconegut, ús gratuït amb restriccions).
  - ChatGPT 4 (versió avançada, de pagament)
  - Google Bard

[13] `eleuther.ai`

Alguns models de llengua existents /3

Per a fer-nos una idea de la grandària dels corpus:

- Una novel·la no massa llarga té uns 100.000 mots.
- Una persona llig uns 200–300 mots per minut.
- Pot, per tant, llegir una novel·la en unes 7 hores.
- *The Pile* té 825 GB de text, uns 50.000.000.000 mots;
- És a dir, uns 500.000 llibres;
- uns 3.500.000 hores = uns 400 anys de lectura continuada.

Índex

Problemàtica i efectes dels models massius de llengua /1

- Entrenar-los (o reentrenar-los) requereix:
  - gestionar i emmagatzemar grans quantitats de text (milers de milions de mots);
  - recursos informàtics especialitzats i molt cars.

  Només unes poques empreses o institucions ho poden fer $\rightarrow$ concentració de poder, dependència.
- Usar-los requereix també ordinadors potents i amb memòria suficient.
- Entrenar-los i usar-los consumeix molta energia, genera gasos d'efecte d'hivernacle i construir-los genera residus tòxics $\rightarrow$ efectes més negatius sobre societats vulnerables.
- Tot i el reentrenament de reforç, poden reproduir biaixos injustos codificats en els textos d'entrenament.

Problemàtica i efectes dels models massius de llengua /2

- Les llengües menors i els grups vulnerables no hi solen estar ben representats:
  - ChatGPT (vers. 24 maig) respon en espanyol o català a preguntes en occità, tot i fer-li-ho notar.
- Produeixen textos nous d'aspecte natural que pareixen escrits per humans.
  - No copien textos del corpus: generen textos nous.
- Les respostes són decebedorament *naturals* però no hi ha garantia que no continguen errors:
  - ChatGPT (vers. 24 maig): "Podríeu llistar alguns dels espais naturals més importants de la ciutat d'Alacant?" $\rightarrow$ "[...] Parc Natural de la Serra de Mariola [...]".

Problemàtica i efectes dels models massius de llengua /3

- És difícil saber, amb la computació distribuïda en milers de neurones i per milions de connexions, com prenen les decisions que els fan generar una resposta.
  - Quan se'ls pregunta per com han raonat per a produir un text, produeixen un text aparentment introspectiu però il·lusori i desconnectat del seu funcionament real.
    - Tot i que probablement basat en el context immediatament anterior.

Problemàtica i efectes dels models massius de llengua /4

- Els models massius de llengua canviaran els mons professional i acadèmic, perquè simplifiquen la generació de textos.
  - Poden reformular, expandir amb una determinada intenció, o resumir.
  - Poden crear esborranys de textos amb intenció literària o artística (per exemple, una estrofa de quatre versos sobre un tema).
  - Poden escriure esborranys de programes d'ordinador en molts llenguatges de programació.
  - Poden preparar qüestions d'examen a partir d'un temari.
  - Poden ajudar a qualificar textos de l'alumnat quan se li donen models i guies.

Índex

Comentaris finals /1

- Els models neurals de llengua generen text que:
  - continua el text anterior, o
  - genera text en resposta a una petició.
- Són *neurals*: es basen a simular el comportament de grans xarxes de neurones artificials inspirades en les dels éssers vius.
- Un model de llengua massiu:
  - es *preentrena* amb grans quantitats de text i de manera no supervisada (predicció del mot següent);
  - es *refina* amb exemples de tasques textuals (de manera supervisada, anotats amb etiquetes);
  - es *refina* encara més amb retroalimentació humana.

  Com a resultat, les respostes mostren un comportament aparentment intel·ligent.

Comentaris finals /2

- Els models massius de llengua com ChatGPT:
  - No tothom els pot crear (calen recursos massius) o executar (calen recursos potents).
  - No són transparents i poden ser injustos.
  - Canviaran molts àmbits professionals i acadèmics relacionats amb la creació i processament de textos.

Aquestes transparències són lliures

Aquest treball es pot distribuir segons els términs de

- la llicència *Creative Commons Reconeixement-CompartirIgual 4.0 Internacional* (`http://creativecommons.org/licenses/by-sa/4.0/deed.ca`)

o

- la Llicència General Pública v. 3.0 de GNU: (`https://www.gnu.org/licenses/gpl-3.0.ca.html`).

Llicència dual! Escriviu-me si voleu el codi font LaTeX: `mlf@ua.es`

Moltes gràcies!

Escanegeu aquest codi QR per a aconseguir una còpia de les diapositives!

O aneu a **`https://t.ly/MkUAp`**.